Out[1]:

Out[2]: Click here to toggle on/off the Python code.

PHYS 10792: Introduction to Data Science

2019-2020 Academic Year

Course instructors: [Rene Breton (http://www.renebreton.org)](http://www.renebreton.org) - Twitter [@BretonRene (https://twitter.com/BretonRene)](https://twitter.com/BretonRene)
[Marco Gersabeck (http://www.hep.manchester.ac.uk/u/gersabec)](http://www.hep.manchester.ac.uk/u/gersabec) - Twitter [@MarcoGersabeck (https://twitter.com/MarcoGersabeck)](https://twitter.com/MarcoGersabeck)

Chapter 9

Syllabus

1. Probabilities and interpretations
2. Probability distributions
3. Parameter estimation
4. Maximum likelihood + extended maximum likelihood
5. Least square, chi2, correlations
6. Monte Carlo basics
7. Probability
8. Hypothesis testing
9. **Confidence level**
10. Goodness of fit tests
11. Limit setting
12. Introduction to multivariate analysis techniques

Topics

9 Confidence level

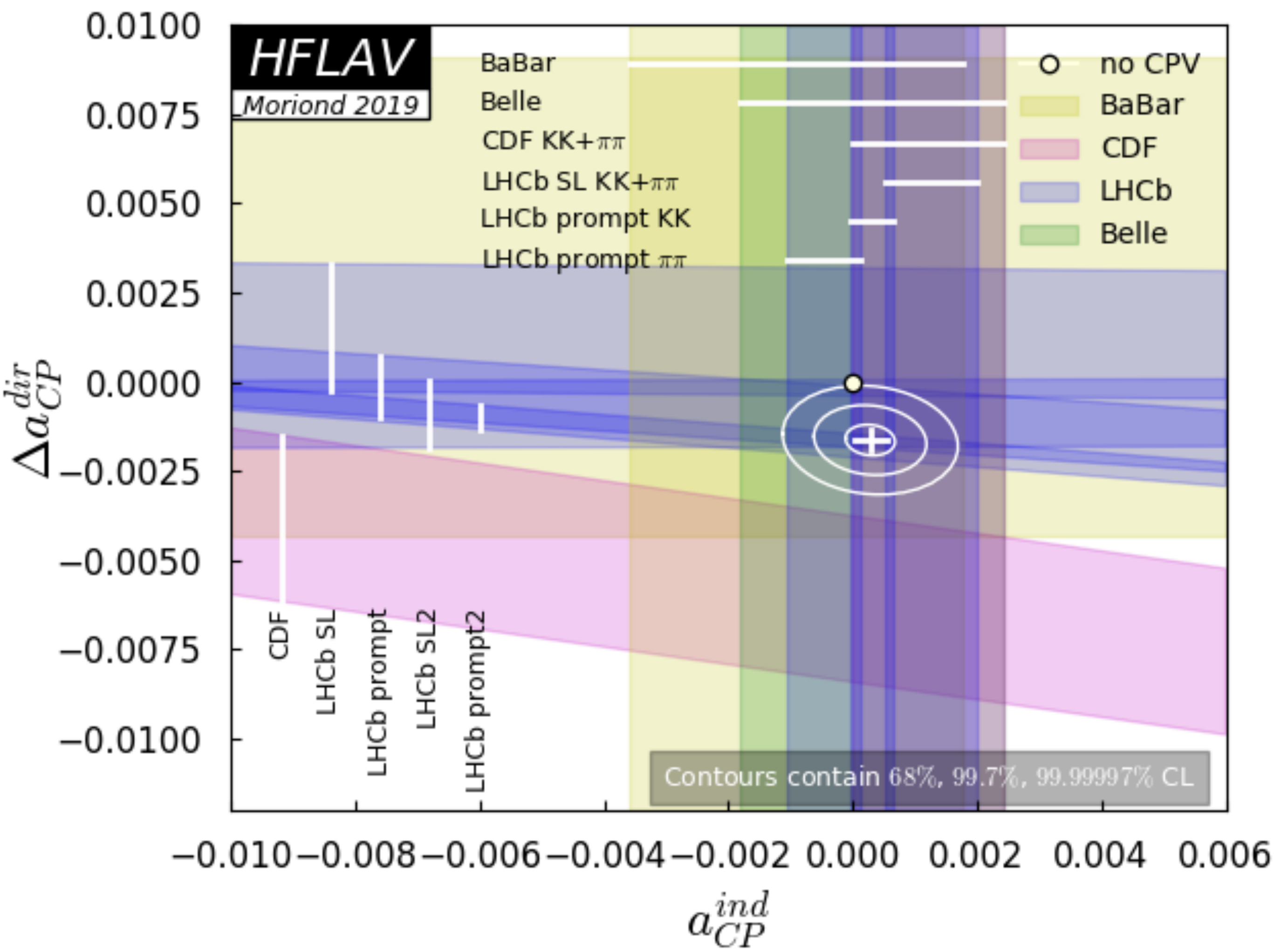
9.1 Confidence intervals in estimation

- 9.1.1 Measuring the impossible
- 9.1.2 Confidence belt construction
- 9.1.3 Measurement of a constrained quantity

9.2 Examples of confidence intervals

- 9.2.1 Binomial confidence intervals
- 9.2.2 Poisson confidence intervals
- 9.2.3 Confidence intervals for several variables: confidence regions

Confidence levels are a central tool to interpreting results such as the discovery of CP violation in charm meson decays.



9 Confidence level

9.1 Confidence intervals in estimation

Suppose we want to know the true value of a parameter μ, having made a measurement with result x and resolution σ. One might be tempted to then say 'μ lies within $x-\sigma$ and $x+\sigma$ with 68% confidence and within $x-2\sigma$ and $x+2\sigma$ with 95% confidence. As described in Sec. 7.5, this is very dangerous and not a sound way of describing the situation.

9.1.1 Measuring the impossible

<u>**Example**</u>

We measure the weight of an empty dish to be 25.30 ± 0.14 g. Adding a sample of powder and repeating the measurement yields 25.50 ± 0.14 g, leading, via subtraction and addition of uncertainties in quadrature, to a measurement of the weight of the powder of 0.2 ± 0.2 g. So far, so valid.

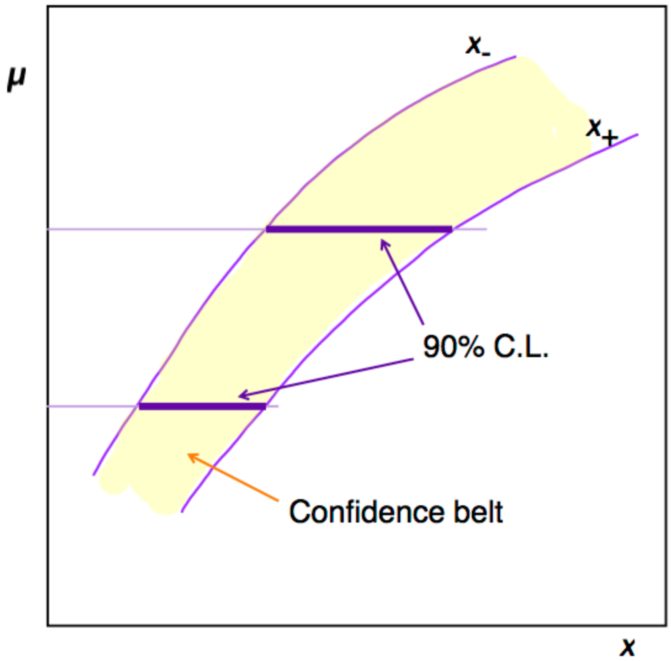
One could now be tempted to say that there is a 68% chance of the actual weight being between 0 g and 0.4 g. But saying that there is a 32% chance of the weight being outside this interval and consequently that there is a 16% chance of the weight being negative is clearly nonesense.

9.1.2 Confidence belt construction

The solution to this issue lies in the construction of confidence belts. To do this we approach the problem from the opposite and maybe less intuitive direction. For a given value of the parameter μ there is a probability distribution function for the measured value x: $P(x;\mu)$. When describing a conventional measurement, this can be a Gaussian function with mean μ and the standard deviation equal to the measurement resolution σ. For a counting experiment it would be a Poisson distribution with mean μ. In general, it will likely peak near $x=\mu$ and fall off to either side.

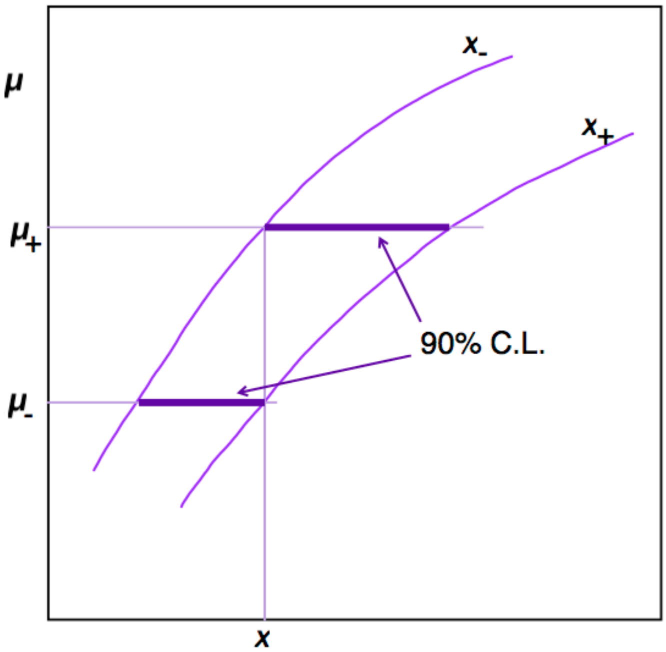
Having this function, we can construct a confidence interval and for the sake of illustration we assume a 90% central interval. This interval means that for a given value of the parameter μ the measured value x will lie with 90% probability in the interval from x_- to x_+. Or, equivalently, for many measurements of μ, 90% of the measured values will lie in said interval.

For different values of μ, we will obtain different values of x_- and x_+. Hence $x_\pm$ can be considered functions of μ. The following figure shows an illustration of this situation. The region between the two curves is called confidence belt. For each value of μ it has a horizontal coverage of 90% confidence level.



The question now becomes, how to relate this to a given measurement outcome x. All of the previous figure can be constructed prior to making a measurement, provided we know the probability density function, e.g. through knowledge of the experimental resolution. This was all based on a horizontal construction. Having measured a particular value x, we can add a vertical line at the measured value. This identifies two points on the vertical parameter axis through the intersections of the line of the measured value with the boundaries of the confidence belt.

The intersection with the x_- curve gives a value of μ for which x is the appropriate x_-. This is the desired upper limit μ_+. This means that if the real value of μ is greater than or equal to μ_+, then the probability of getting a measurement smaller than this is 5% or less. Analogously, we can deduce μ_- from the intersection with the x_+ curve. We therefore quote the 90% confidence interval for the true value μ as the range μ_- to μ_+ as shown in the following figure.



Interpretation

We can now make the statement that the true value of μ lies between μ_- and μ_+ with 90% probability. Note that this is only a statement about the boundaries μ_- and μ_+ and not about μ itself!

A given statement from such a construction at a confidence level of say 90%, e.g. the electron mass lies between 510 and $515\,\mathrm{keV}/c^2$, is either right or wrong. However, taking a large number of such statements, 90% of them will be true. This provides a powerful test of the reliability of uncertainties quoted in different scientific communities. Some particle physics experiments for example produce hundreds of measurements throughout their lifetime. A good number of them either test previous measurements at greater precision or are themselves verified by subsequent updates. Many are quoted with uncertainties representing a standard deviation of 68% confidence level. Hence, less than 5% of all measurements should differ from the true value by more than two standard deviations. It would make for a very interesting project to check this.

Confidence levels from Gaussians

In the case of Gaussian distribution functions, the construction becomes very simple. The x_- and x_+ curves become straight lines and the limits are obtained simply by $\mu_\pm = x \pm n\sigma$, where $n = 1$ for 68% confidence level, $n = 1.64$ for 90% confidence level, and so on.

The following code produces some of the most common numbers.

```
sigma | C.L.
------------
    1 | 0.68
    2 | 0.954
    3 | 0.9973
    4 | 0.999937
    5 | 0.99999943

C.L. | sigma
------------
0.90 | 1.64
0.95 | 1.96
```

9.1.3 Measurement of a constrained quantity

The measurement of a mass was mentioned in the introduction to confidence levels. The construction with a confidence belt fails miserably when a negative mass is measured with a relatively small uncertainty, e.g. (-0.5 ± 0.2) g. When trying to construct a 90% confidence interval, we would get $\mu_\pm = (-0.5 \pm 0.2 \times 1.64)$ g, i.e. even $\mu_+ = -0.172$ g remains negative.

A solution to this issue is a Bayesian construction with a normalisation that takes the physical limit into account. For example for positive masses, one gets

$$P(\mu|x) = \frac{e^{-(x-\mu)^2/2\sigma^2}}{\int_0^\infty e^{-(x-\mu')^2/2\sigma^2}\,d\mu'}(x < 0).$$

This construction will then lead to one limit being zero, i.e. we set an upper limit. We will discuss this in more detail in a few weeks within the topic of limit setting.

9.2 Examples of confidence intervals

9.2.1 Binomial confidence intervals

For Binomial distributions, events belong to exactly one of two classes, e.g. true or false, greater or smaller than a threshold, male or female, etc. This applies to samples of finite size and so the observed events are discretely distributed. Contrary to that the true variable is continuous, e.g. the probability for an event to be true can take any value and even the expectation value, i.e. the probability multiplied by the sample size is not necessarily an integer.

Recall the Binomial distribution

$$P(k; p, n) = p^k (1-p)^{n-k} \frac{n!}{k!(n-k)!},$$

where p is the probability of success, n is the sample size, and k is the number of successes for this sample.

The expectation value is

$$< k >= np.$$

Construction of a Binomial confidence belt

Given that the distribution of events is discrete, the integrals used in the construction of confidence intervals have to be replaced by sums. Recall that for a central interval, for a given confidence level C, we have to determine

$$\int_{-\infty}^{x_-} P(x)dx = \int_{x_+}^{\infty} P(x)dx = (1-C)/2.$$

The direct replacement would lead, for a given confidence level C, to

$$\sum_{k=0}^{k_-} P(k; p, n) = \sum_{k=k_+}^{n} P(k; p, n) = (1-C)/2.$$

In general the discrete nature of k will prevent these equalities to be satisfied exactly. Therefore they have to be replaced by inequalities that guarantee that the confidence interval covered by the range k_- to k_+ is *at least* C. This is given by the following constructions

$$\sum_{k=0}^{k_+} P(k; p, n) \geq 1 - (1-C)/2.$$

and

$$\sum_{k=k_-}^{n} P(k; p, n) \geq 1 - (1-C)/2.$$

If we are to construct bands with $C = 0.9$, these two equations mean that we have to construct one-sided intervals that each cover at least 0.95. Their intersection, i.e. the range k_- to k_+ will then cover at least 0.9.

Finally, if m successes are observed, the limits on the true probability interval can be assigned with p_- and p_+ given by

$$\sum_{k=m+1}^{n} P(k; p_+, n) = 1 - (1-C)/2,$$

and

$$\sum_{k=0}^{m-1} P(k; p_-, n) = 1 - (1 - C)/2.$$

In practice, these are the outward-facing corners of the confidence belt at a position $k = m$. These are also known as the *Clopper-Pearson confidence limits*.

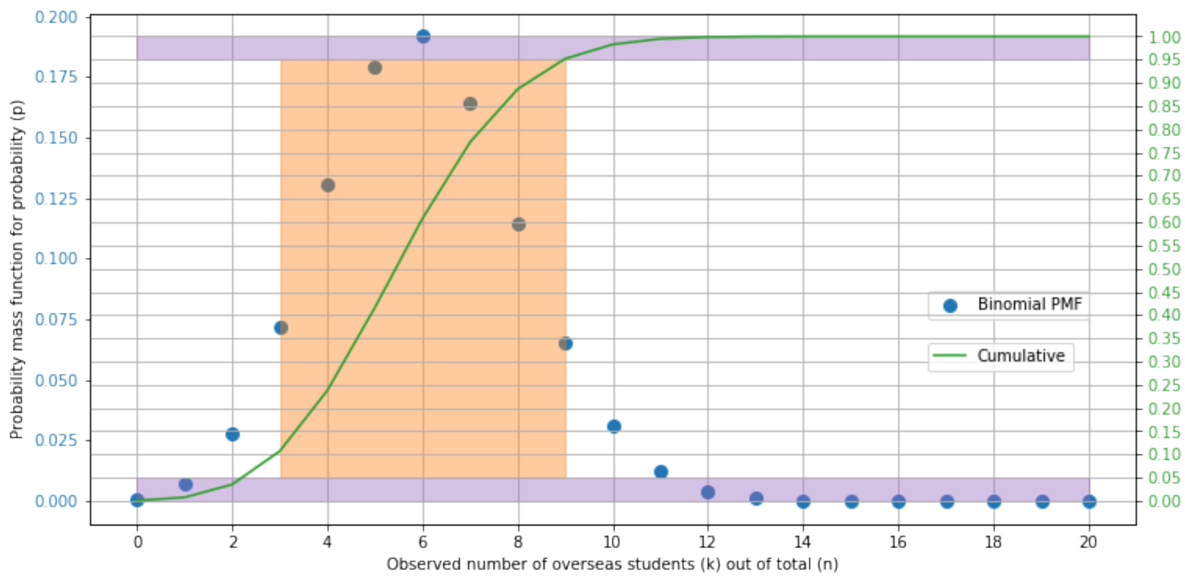
Example: Fraction of overseas students

Let's first establish the number of overseas students in the course.

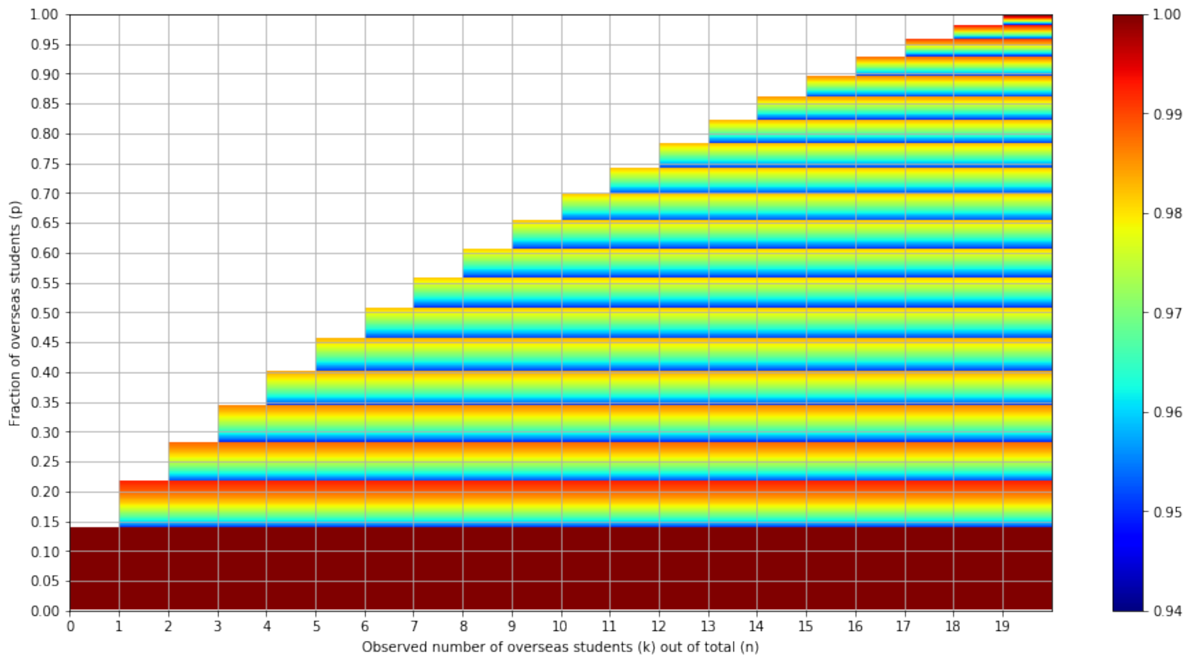
- Over to [menti.com (https://www.menti.com)](https://www.menti.com) with code 88 12 11

Construction of a confidence belt based on these data

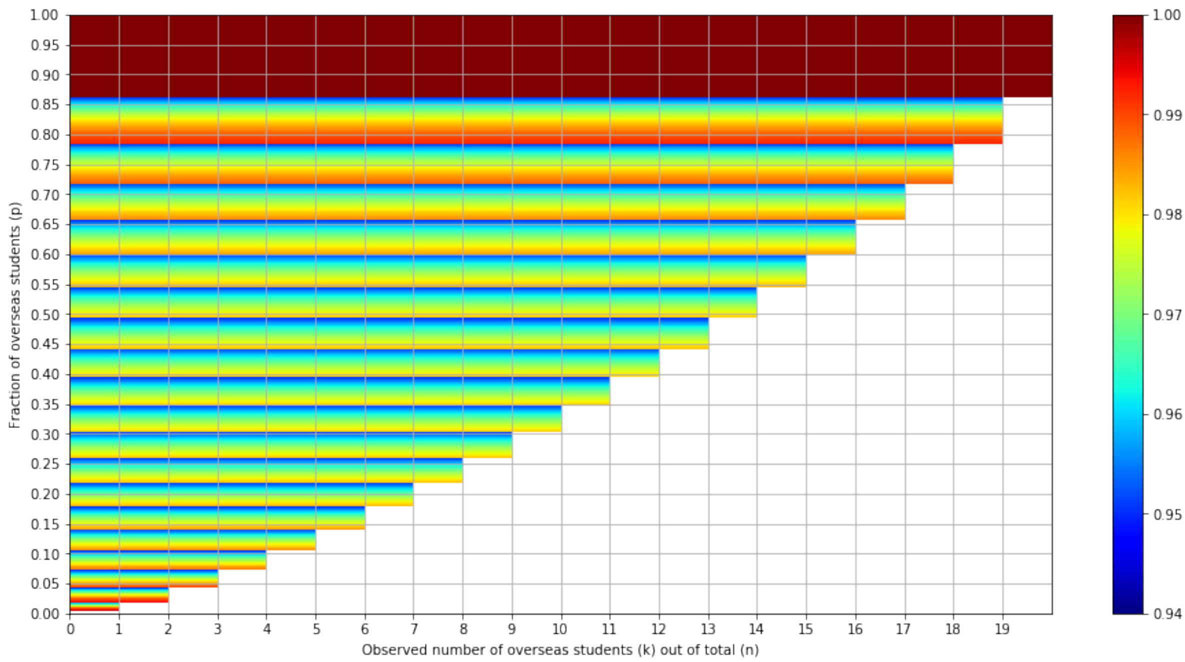
At first, let's just look at the construction for one given value of p.



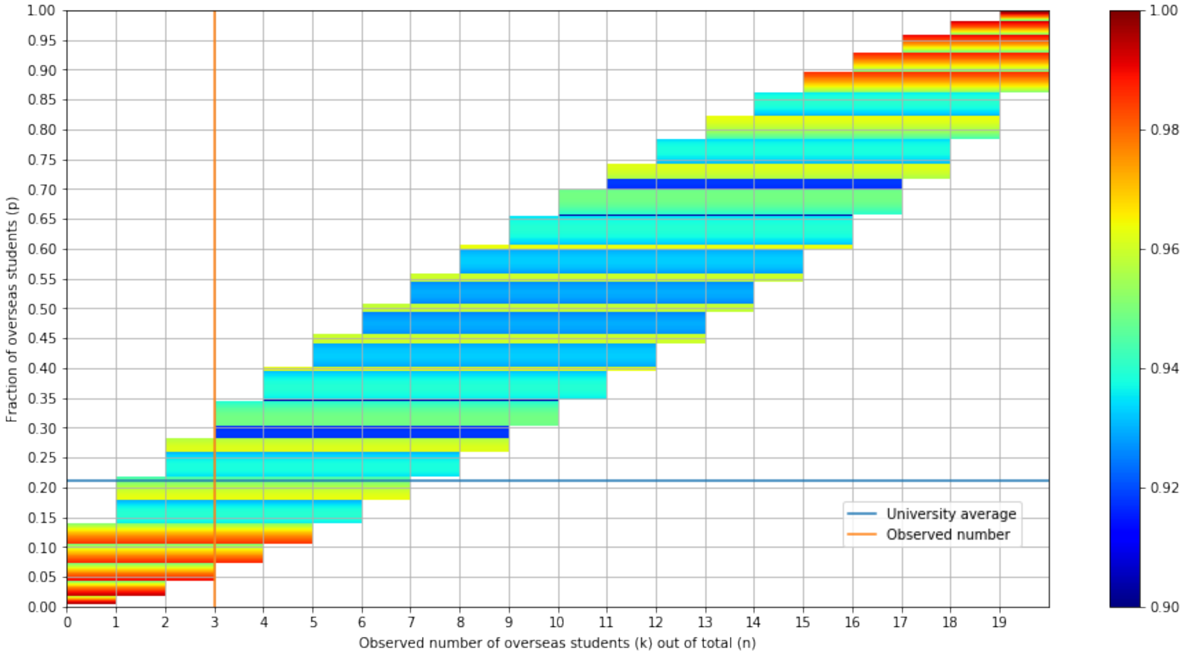
The following three plots show the construction of the confidence belt first separated in one-sided intervals, which satisfy the inequalities discussed above, and finally their intersection. The color shows the variation of the confidence level covered, which illustrates the steps.



The plot above can be read from bottom to top, i.e. with increasing p, as the lower edge of the interval stepping to the right as soon as the resulting coverage exceeds 0.95. Therefore, it always reaches 0.95 at the bottom of a band of the same k.



The plot above can be read from bottom to top, i.e. with increasing p, as the upper edge of the interval stepping to the right as soon as the resulting coverage would otherwise drop below 0.95. Therefore, it always reaches 0.95 at the top of a band of the same k.


Fraction of overseas students (p)
Observed number of overseas students (k) out of total (n)
University average
Observed number
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
1.00
0.98
0.96
0.94
0.92
0.90

9.2.2 Poisson confidence intervals

The procedure for a Poisson process is rather similar to that of a Binomial distribution with the main difference that the range of possible values is not limited by a total number but goes to infinity (this is exactly how the distribution is defined, by taking the $n \to \infty$ limit of the Binomial distribution function).

Recall the Poisson distribution

For a given number of observed events k and an expectation value of λ, the Poisson probability is given by

$$P(k; \lambda) = \frac{e^{-\lambda}\lambda^k}{k!}.$$

Construction of Poisson intervals

To construct central 90% intervals, we need the greatest value of N_- that satisfies for a given λ

$$\sum_{k=N_-}^{\infty} P(k; \lambda) \geq 0.95.$$

This is equivalent to

$$\sum_{k=0}^{N_- -1} P(k; \lambda) \leq 1 - 0.95 = 0.05,$$

which is easier to calculate.

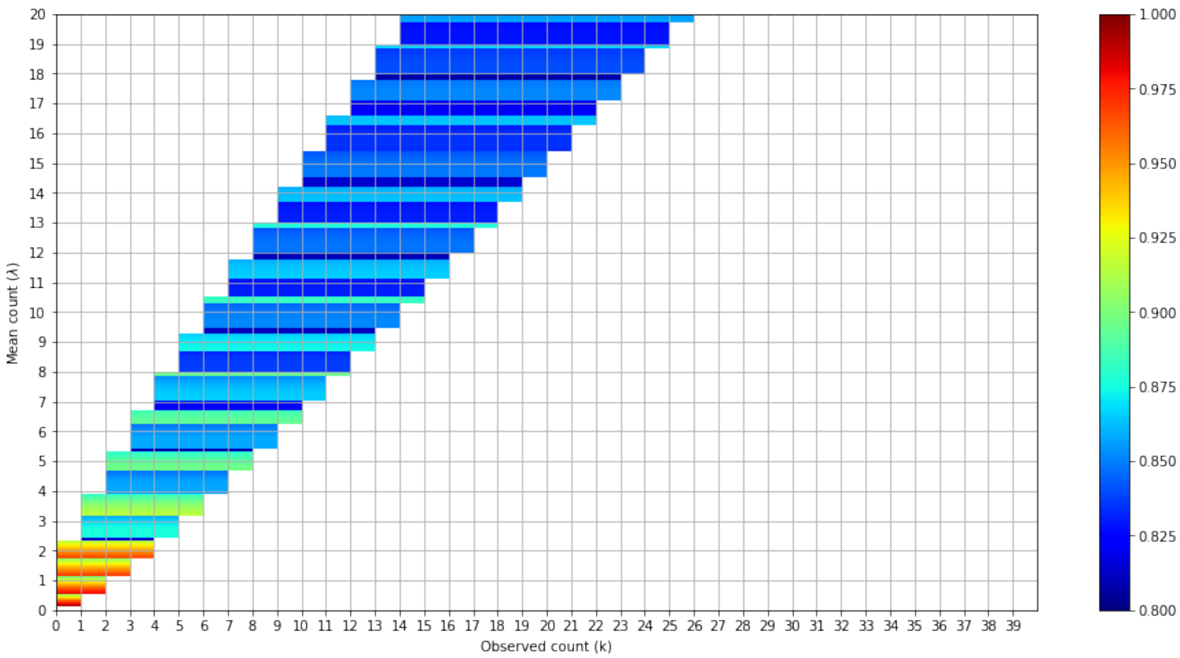
Accordingly, we require the smallest N_+ that satisfies for a given λ

$$\sum_{k=0}^{N_+} P(k; \lambda) \geq 0.95.$$

Practical example

One useful number to remember is the 90% upper limit for the case that the number of observed events is 0; this limit is $\lambda_+ = 2.3$. Hence, in any counting experiment that yields an obervation of 0, we can be 90% sure that the true number of events is no greater than 2.3.

The 90% upper limit corresponds to the left-hand side of an 80% central-interval belt. This is shown below where the 2.3 can be read off as the top of the belt at its left-hand end at $k = 0$.



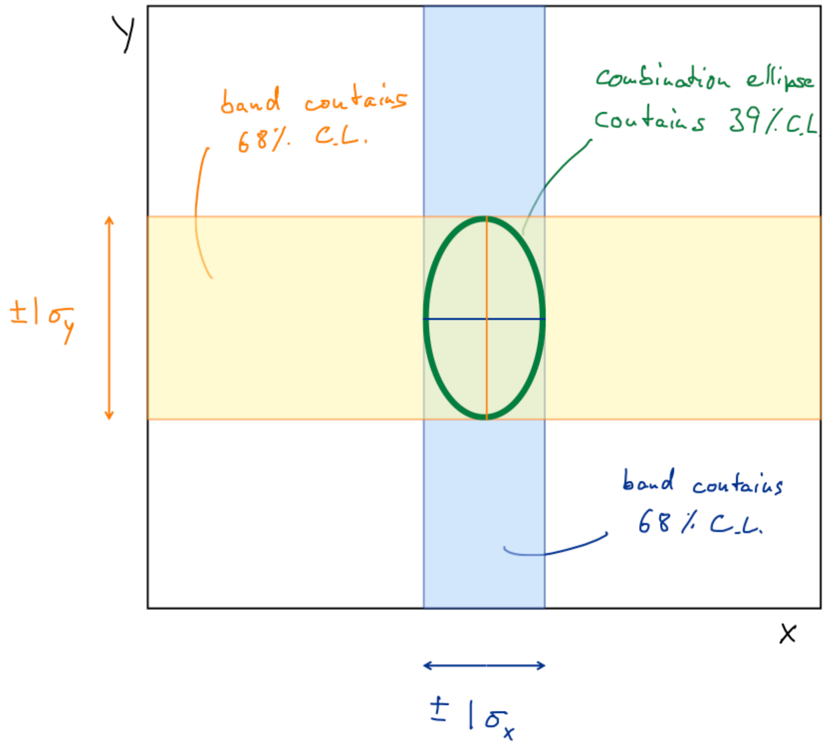
9.2.3 Confidence intervals for several variables: confidence regions

In the case of several variables the confidence intervals does not apply directly, e.g. in the case of two variables, the parameter space corresponding to a certain confidence level is not given by independent intervals in the individual variables, which would result in a rectangular area in parameter space, but rather by a more complex confidence region.

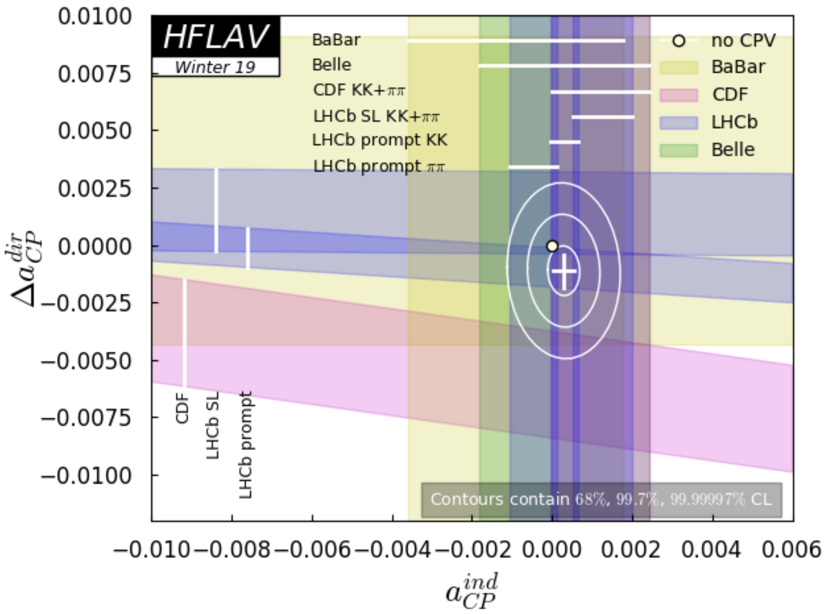
In the case of Gaussian uncertainties, the confidence region in two dimensions is an ellipse. However, this can take more complicated shapes depending on how the parameters are determined and on whether Gaussian uncertainties apply.

In general, the maximum likelihood method provides a straghtforward way of constructing confidence regions as the likelihood is multiplicative and hence the likelihood defining the confidence region is the product of the likelihoods of the individual vairables, potentially accounting for correlations.

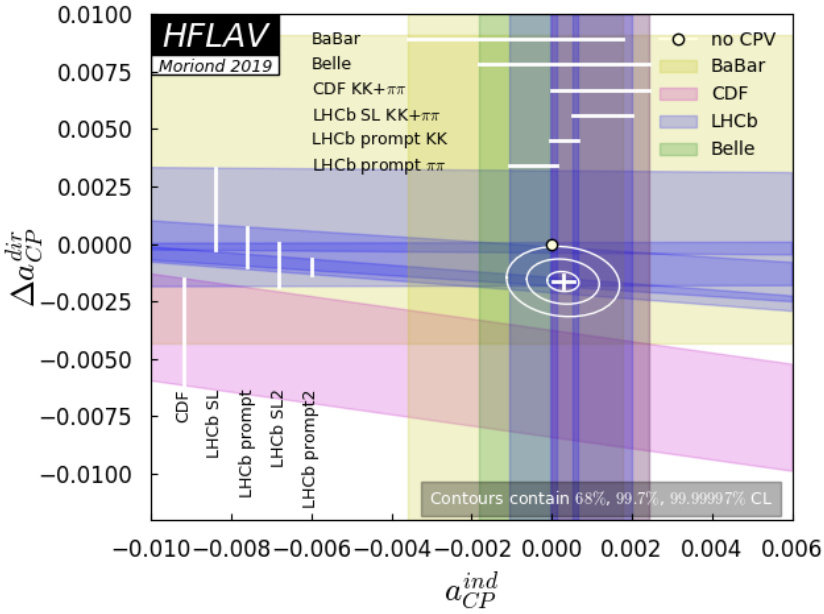
One caveat applies regarding the confidence level of a multi-dimensional region compared to a one-dimensional interval. As illustrated below, the ellipse spanned by the 1D 68% confidence-level intervals, i.e. the standard deviations of the measurements, contains only 39% confidence level as the remainder is covered in infinitely long bands limited only in one variable. It is therefore essential to state the coverage of any multi-dimensional confidence region to avoid confusion and mistakes.



Below are a few real-life examples of more comlicated confidence contours. These are taken from my work in the [Heavy Flavor Averaging Group (https://hflav.web.cern.ch)](https://hflav.web.cern.ch) and more details for those who want it can be found on the linked web page, where these particular examples refer to *charm physics*.



And as of a year ago (see e.g. [the LHCb publication (http://inspirehep.net/record/1726338)](http://inspirehep.net/record/1726338) or an article in [The Conversation (https://theconversation.com/cern-study-sheds-light-on-one-of-physics-biggest-mysteries-why-theres-more-matter-than-antimatter-113947)](https://theconversation.com/cern-study-sheds-light-on-one-of-physics-biggest-mysteries-why-theres-more-matter-than-antimatter-113947)):



Note: The content of this Jupyter Notebook is provided for educational purposes only.